MEMORIA
PROYECTO USISA

Amalio Gómez López

1-Introducción

El proyecto trata acerca de datos de una empresa llamada USISA (Unión Salazonera Isleña S. A.). Ubicada en Isla Cristina (Huelva) y con más de 40 años en la industria, es la conservera más grande de Andalucía.

1.1-Motivación

La empresa tiene como objetivo mejorar los números de venta de su tienda online. Para ello se propone ofrecer una campaña de ofertas a un grupo concreto de clientes, porque no es viable en cuanto a recursos alcanzar al total de clientes.

Diferenciar y agrupar los clientes es la piedra angular del proyecto. Se trata de caracterizar a estos clientes por sus hábitos de compra. Se tiene en cuenta la frecuencia de compra, lo reciente que han comprado y el dinero que han gastado.

2-Objetivos

2.1- Principal

El proyecto busca dar respuesta a la pregunta anteriormente desarrollada para concluir el grupo óptimo de clientes (así como un modelo que los identifique) para ofrecer una campaña de ofertas. Los grupos tienen que estar correctamente definidos y tiene que tener sentido las variables que caracterizan a cada uno.

La hipótesis inicial es que, siguiendo el sentido común, los clientes más habituales que realizan muchos pedidos y gastan mucho dinero serían más interesantes para enfocar la campaña. Al contrario que a un grupo de clientes perdido del que se puede recuperar un porcentaje.

2.2-Secundarios

Uno de los objetivos secundarios es la agrupación de productos. Comprobar las combinaciones de productos que realizan los clientes por pedido individual. El propósito de esto es ver si existe algún patrón en la compra.

El segundo objetivo secundario es un sistema de recomendación que se sugiere para implementar la campaña de ofertas. En lugar de ofrecer descuentos directos sobre el precio final en tienda. Se sugiere ofrecer descuentos en productos específicos que el algoritmo sugiere a los clientes. El objetivo de esto sería incentivar aún más la compra al mismo tiempo que se recuperan clientes o se establece un programa de fidelización en función del resultado del clustering en el objetivo principal.

3- Datos

Los datos proporcionados por la empresa están divididos en dos bases de datos. El primero consiste en la base de datos antigua de la tienda online desde el 2017 hasta principios de 2022. La nueva comprende desde principios de 2022 hasta la actualidad.

El presente proyecto emplea los datos de la base nueva para la segmentación de clientes y demás modelos. El histórico con la base de datos antigua se utiliza solo a modo de visualización.

3.1- Preprocesamiento y limpieza

Una de las grandes labores de este proyecto ha sido la limpieza y preprocesamiento de datos. Dado que la empresa no utiliza los datos que recaba para nada, no han reparado en muchos errores de recogida de los mismos.

Se ha escogido la información relevante como tipo de producto, identificación de pedidos y de clientes, códigos postales y demás. Se han estandarizado los datos y se han pasado a variables numéricas. Asimismo, se han eliminado los pedidos de prueba y duplicados.

4-Análisis exploratorio de los datos

Se realiza un análisis exploratorio en el que se grafica el mapa con la localización de los pedidos según el código postal del pedido. Además, se grafican las variables que vamos a usar para nuestros modelos de segmentación como dinero total, cantidad, tipo de productos, entre otros. Se usa un mapa de calor para ver las correlaciones hallando que solo existe en tipo de marca y tipo de aceite, así como peso y precio que van de la mano. Se comprueba la distribución de las variables.

Gráficos de timeseries

Por otro lado, se realizan las series temporales sobre el número de pedidos y el dinero factura tanto mensualmente como semanalmente.

Además a modo curiosidad, se realiza un modelo *SARIMAX* de predicción a futuro de pedidos y dinero facturado. Antes se prueba y se ajustan los parámetros para los últimos 3 meses, y se usan las métricas *MSE*, y *MAE*.

5- Segmentación de clientes

5.1- Clustering con el modelo KMeans

Se compara con el modelo *DBSCAN*, se optimizan parámetros para el mismo. Se comprueba el *silhouette score* para ver qué modelo define mejor los clusters.

Elbow method para comprobar el número de clusters óptimo y se realizan varias ejecuciones de prueba.

Se usa la etiqueta de los clusters con otros modelos como *RandomForestClassifier*, *LogisticRegression* y *XGBClassifier* para ver que tal predicen y explican la etiqueta arrojando una media del 98% de precisión.

5.2-Explicación de los clusters

Se agrupan los clientes por sus respectivos clusters y se obtiene el centro de cada cluster usando la media para ver las características que mejor explican cada grupo.

-Primer clúster: se observa que la última compra de este grupo ha sido muy reciente, que realiza pedidos habitualmente y que gasta moderadamente. Comprende el 15% aproximadamente del total de clientes.

-Segundo clúster: su última compra fue hace mucho, promedian tan solo un pedido y es el grupo que menos dinero gasta. Corresponde al 63% de los clientes.
-Tercer clúster: su última compra ha sido hace relativamente poco, es el grupo que más pedidos promedian y que más gasta prácticamente cuadruplicando los demás. Se trata del 22% de los clientes.

6-Conclusiones
El objetivo principal se logra de manera eficiente. Se distinguen los diferentes grupos de clientes además de explicar sus características. Cada cliente se identifica correctamente.

En cuanto a la hipótesis, se confirma que el grupo de cliente óptimo al que dedicar la campaña de ofertas es el segundo y último clúster en el que los clientes gastan mucho dinero y compran a menudo (al por mayor).

Objetivo secundarios, se logran de manera parcial el sistema de recomendación y el clustering por productos al carecer de las valoraciones de los clientes como se comentará en el siguiente párrafo.

Por otro lado una de las conclusiones principales es que pocos clientes se mantienen y compran frecuentemente, lleva a pensar que o se registran erróneamente o se van.

7- Limitaciones y áreas de mejora

Una de las limitaciones del proyecto ha resultado ser el tiempo. Se tuvo que dedicar mucho tiempo al preprocesamiento de datos. Además, el proyecto abarca puntos diferentes que se pueden mejorar ampliamente. También se podría mejorar el sistema de recomendación si mejoramos la estandarización de productos y se añaden valoraciones. Esto también ha dificultado el clustering por productos dentro de un pedido para ver si existía un patrón en las combinaciones.

Hay algunos puntos que mejorarían mucho la base de datos y los mismos para poder desarrollar futuros proyectos y su desempeño. El tener los productos no estandarizados ha resultado costoso para la limpieza de datos y dificulta los modelos de cluster. A modo de idea sería positivo guardar los datos independientemente del peso o del empaquetado y sean tratados como el mismo producto. Un punto a mejorar también sería el idioma de la recogida de datos porque había pedidos del extranjero recogidos en inglés.

Otro punto a mejorar podría ser el introducir un sistema de valoración de productos comprados y reseñas dentro de la tienda online y que se recojan en la base de datos. Esto habría hecho posible mejorar enormemente el sistema de recomendación desarrollado con este limitante en el presente proyecto.

8- Futuras líneas de investigación para proyectos
Tras la estandarización de productos una buena línea para un futuro proyecto sería el *churn* que analice esa rotación de clientes. Además sería interesante analizar los productos en carrito que no se llegan a comprar, para lo cual sería necesario que se recogieran estos datos en la base.